## Misuse of RPKM or TPM normalization when comparing across samples and **sequencing protocols**<br>Shanrong Zhao<sup>1\*</sup>, Zhan Ye<sup>2</sup>, Robert Stanton<sup>11</sup> Integrative Biology Center of Excellence

 $\frac{1}{2}$ Shanrong Zhao", Zhan Ye<sup>-</sup><br><sup>2</sup> Early Clinical Developmer<br>Pfizer Worldwide Research<br>\* Corresponding author: Sl , Robert Stanton<sup>--</sup> Integrative Biology Center or Excellence<br>nt<br>n and Development, Cambridge, MA 02139, USA<br>hanrong.Zhao@pfizer.com Pfizer Worldwide Research and Development, Cambridge, MA 02139, USA<br>\* Corresponding author: Shanrong.Zhao@pfizer.com

Phase Presearch and Development, Cambridge, MA 12117, 1988<br>\* Corresponding author: Shanrong.Zhao@pfizer.com<br>Abstract

# $\ddot{\phantom{0}}$ Abstract

In recent years RNA-sequencing (RNA-seq) has emerged as a powerful technology for transcriptome |
|
|
| |
|
|
| In recent years into experiencing (introduct) has emerges as a powerful technology for transcriptions<br>profiling. For a given gene, the number of mapped reads is not only dependent on its expression level<br>and gene length, b proming. For a given gene, the number of mapped reads in the only dependencies, RPKM (Reads Per<br>Kilobase of transcript per Million reads mapped) and TPM (Transcripts Per Million) are used to measure<br>gene or transcript expr Rilobase of transcript per Million reads mapped) and TPM (Transcripts Per Million) are used to measure<br>gene or transcript expression levels. A common misconception is that RPKM and TPM values are already<br>normalized, and th Represent the relative should be comparable across samples or RNA-seq projects. However, RPKM and TPM represent the relative abundance of a transcript among a population of sequenced transcripts, and therefore depend on th mormalized, and thus should be comparable across samples or RNA-seq projects. However, RPKM and<br>TPM represent the relative abundance of a transcript among a population of sequenced transcripts, and<br>therefore depend on the TPM represent the relative abundance of a transcript among a population of sequenced transcripts, and<br>therefore depend on the composition of the RNA population in a sample. Quite often, it is reasonable to<br>assume that tota Therefore depend on the composition of the RNA population in a sample. Quite often, it is reasonable to assume that total RNA concentration and distributions is very close across compared samples.<br>Nevertheless, the sequenc therefore depends to the component on the component of the RNA concentration and distributions is very close across compared samples.<br>Nevertheless, the sequenced RNA repertoires may differ significantly under different exp Nevertheless, the sequenced RNA repertoires may differ significantly under different experimental<br>conditions and/or across sequencing protocols; thus, the proportion of gene expression is not directly<br>comparable in such ca Conditions and/or across sequencing protocols; thus, the proportion of gene expression is not directly<br>comparable in such cases. In this review, we illustrate typical scenarios in which RPKM and TPM are<br>misused, unintentio comparable in such cases. In this review, we illustrate typical scenarios in which RPKM and TPM are<br>misused, unintentionally, and hope to raise scientists' awareness of this issue when comparing them<br>across samples or diff misused, unintentionally, and hope to raise scientists' awareness of this issue when comparing them misused, unintentionally, and hope to raise scientific and hold is this issue when comparing them<br>across samples or different sequencing protocols.<br>Keywords: RNA-seq, normalization, RPKM, FPKM, TPM

Keywords: RNA-seq, normalization, RPKM, FPKM, TPM  $\overline{a}$ Keywords: RNA-seq, normalization, RPKM, FPKM, TPM

### Introduction

In recent years, RNA-seq has emerged as a powerful technology for transcriptome profiling (Mortazavi et al. 2008; Zhao et al. 2014; Zhao et al. 2015). In 2008, Mortazavi et al. used RNA-seq to quantify transcript prevalenc transcript prevalence for the first time (Mortazavi et al. 2008). RNA-seq avoids some of the technical<br>limitations of microarrays, including varying probe performance, cross-hybridization, nonspecific<br>hybridization, and dy limitations of microarrays, including varying probe performance, cross-hybridization, nonspecific minimization, and dynamic range issues. RNA-seq can also detect low abundance transcripts, novel<br>transcripts, alternative splice forms of transcripts, genetic variants and gene fusions (Zhao et al. 2014;<br>Zhang et al. 2018) Transcripts, alternative splice forms of transcripts, genetic variants and gene fusions (Zhao et al. 2014;<br>Zhang et al. 2018). Because RNA-seq does not rely on a pre-designed complementary sequence<br>detection probe, it is n Thang et al. 2018). Because RNA-seq does not rely on a pre-designed complementary sequence<br>detection probe, it is not limited to the interrogation of selected probes on an array and can also be<br>applied to species for which detection probe, it is not limited to the interrogation of selected probes on an array and can also be<br>applied to species for which the whole reference genome is not yet assembled. Thus, RNA-seq delivers<br>both less biased a applied to species for which the whole reference genome is not yet assembled. Thus, RNA-seq delivers<br>both less biased and previously unknown information about the transcriptome.<br>In a standard RNA-seq experiment, RNAs from

In a standard RNA-seq experiment, RNAs from different sources (blood, tissue, cell lines) are purified, both less biased and previously unknown interesting and the transcriptome.<br>In a standard RNA-seq experiment, RNAs from different sources (blood, tissue<br>typically enriched with oligo (dT) primers, and then fragmented. After In a state of the state of the state of the state selection, millions or even<br>In a state selection, millions or even<br>In a state selection, millions or even<br>In a state selection of solution of the major steps in RNA-seq dat billions of short sequence reads are generated from a randomly fragmented cDNA library (Zhao et al.<br>2015; Zhao et al. 2018). The major steps in RNA-seq data analysis include quality control, read<br>alignment, quantification billions of al. 2018). The major steps in RNA-seq data analysis include quality control, read<br>alignment, quantification of gene and transcript expression levels, normalization, analysis of differential<br>gene expression, cha 2011; Conesa et al. 2016; Zhao et al. 2016). RNA-seq has a wide variety of applications in biological recession, discovery and davelerment (Khatson et al. 2011; Conesa et al. 2016; Zhao et al. 2016). RNA-seq has a wide var gene expression, characterization of alternative splicing, functional analysis and gene fusion detection.<br>The algorithms and challenges associated with each step have been reviewed elsewhere (Garber et al.<br>2011; Conesa et The algorithms and challenges associated with each step have been reviewed elsewhere (Garber et al.<br>2011; Conesa et al. 2016; Zhao et al. 2016). RNA-seq has a wide variety of applications in biological<br>research, drug disco The algorithms are to all 2016; Zhao et al. 2016). RNA-seq has a wide variety of applications in biological<br>research, drug discovery and development (Khatoon et al. 2014). However, the most common and<br>popular application o 2012) Extracted the ELLY and ELLY and ELLY interesting the a mass cannot propplement in an energy of research, drug discovery and development (Khatoon et al. 2014). However, the most common and popular application of RNA-s research, and development (Mattersearch and development (Mattersearch and development and popular application of RNA-seq is the identification of differentially expressed genes (DEGs) or isoforms<br>between two or more condit popular approximate conditions. These DEGs may serve as drug targets and biomarkers for clinical<br>diagnosis, improve our understanding of disease pathophysiology, help determining a compound's<br>mechanism of action, and assis diagnosis, improve our understanding of disease pathophysiology, help determining a compound's<br>mechanism of action, and assist with patient stratification (Khatoon et al. 2014).<br>Measures of expression: RPKM/FPKM and TPM

## Measures of expression: RPKM/FPKM and TPM

mechanism of action, and assist with patient stratification (Khatoon et al. 2014).<br>
Measures of expression: RPKM/FPKM and TPM<br>
In RNA-seq, the expression level of each mRNA transcript is measured by the total number of map Measures of expression: RPKM/FPKM and TPM<br>In RNA-seq, the expression level of each mRNA transcript is measured by the tragments, which is expected to be directly proportional to its abundance |
|
|
| In Fragments, which is expected to be directly proportional to its abundance level. However, after calculating the read counts, data normalization is essential to ensure accurate inference of gene expressions (Dillies et a Fragments, data normalization is essential to ensure accurate inference of gene<br>expressions (Dillies et al. 2013; Li et al. 2015; Evans et al. 2018). Raw counts mapped to a given gene are<br>not comparable between samples or expressions (Dillies et al. 2013; Li et al. 2015; Evans et al. 2018). Raw counts mapped to a given gene are<br>not comparable between samples or conditions because the sequencing depths or library sizes (the total<br>2 not comparable between samples or conditions because the sequencing depths or library sizes (the total  $n = 2$ 

number of mapped reads) typically vary from sample to sample. Raw counts of different genes within<br>one sample are also not directly comparable, because longer transcripts have more reads mapped to<br>them compared with shorte one sample are are increasely comparable, because longer transcripts have more reade implements<br>them compared with shorter transcripts of a similar expression level. Therefore, instead of using integer<br>counts directly, nor them counts directly, normalized expression units such as RPKM (Reads Per Kilobase of transcript per Million<br>reads mapped), FPKM (Fragments Per Kilobase of transcript per Million fragments mapped), and TPM<br>(Transcripts Per reads mapped), FPKM (Fragments Per Kilobase of transcript per Million fragments mapped), and TPM<br>(Transcripts Per Million), are necessary to remove technical biases in sequenced data. FPKM is closely<br>related to RPKM except Transcripts Per Million), are necessary to remove technical biases in sequenced data. FPKM is closely related to RPKM except with fragment (a pair of reads) replacing read (the reason for this nomenclature is historical, s related to RPKM except with fragment (a pair of reads) replacing read (the reason for this nomenclature<br>is historical, since initially reads were single-end, but with the advent of paired-end sequencing it now<br>makes more s related to RPK and the PPL with registers (a pair of reads) replacing read (the reason for this intentional in<br>
makes more sense to speak of fragments, and hence FPKM).<br>
RPKM was initially introduced to facilitate transpar

makes more sense to speak of fragments, and hence FPKM).<br>RPKM was initially introduced to facilitate transparent comparison of transcript levels both within and<br>between samples, as it re-scales gene counts to correct for d makes more in the sense of the RPKM was initially introduced to facilitate transparent compressive to the SPKM of the SPKM (Mortazavi et al. 2008). Since RPKM was introduced, if the speak of the SPAM of the SPAM of the SPA |
|
| Between samples, as it re-scales gene counts to correct for differences in both library sizes and gene<br>length (Mortazavi et al. 2008). Since RPKM was introduced, it has been widely used due to its simplicity.<br>RPKM =  $10^9$ bend in the samples, as it research gene counts to convert for differences in a sum many, sizes and generation<br>
length (Mortazavi et al. 2008). Since RPKM was introduced, it has been widely used due to its simplicity.<br>
RP

$$
RPKM = 109 *  $\frac{Reads \ mapped \ to \ the \ transcript \ length}{Total \ reads \ * \ Transfer \ length}$
$$

 $\text{RPKM} = 10^9 * \frac{\text{Reads mapped to the transcript}}{\text{Total reads} * \text{Transcript length}}$ <br>The intended meaning of RPKM is a measure of <u>r</u>elative RNA <u>m</u>olar concentration (rmc) of a transcript in<br>a sample. If a measure of RNA abundance is proportional to rmc, then the Total reads \* Transcript length<br>re of <u>r</u>elative RNA <u>m</u>olar <u>c</u>oncentrati<br>is proportional to rmc, then their av<br>inverse of the number of transcript: ך<br>}<br>} a sample. If a measure of RNA abundance is proportional to rmc, then their average over genes within a<br>sample should be a constant, namely the inverse of the number of transcripts mapped. Unfortunately,<br>RPKM does not respe sample should be a constant, namely the inverse of the number of transcripts mapped. Unfortunately,<br>RPKM does not respect this invariance property and thus cannot be an accurate measure of rmc<br>(Wagner et al. 2012). In fact Sample to sample to sample to sample to sample show the inverse of the i (Wagner et al. 2012). In fact, the average RPKM varies from sample to sample. Therefore, TPM<br>(Transcripts Per Million), a slight modification of RPKM, was proposed (Li and Dewey 2011; Wagner et al.<br>2012).<br>TPM =  $10^6 * \frac{reads$ (Transcripts Per Million), a slight modification of RPKM, was proposed (Li and Dewey 2011; Wagner et al.<br>2012).<br>IPM =  $10^6 * \frac{reads\ mapped\ to\ transcript\ /\ transient\ length}$ <br>TPM and RPKM are closely related. It is straightforward to convert a RPKM to

TPM = 
$$
10^6 * \frac{reads \ mapped \ to \ transcript \ /\ transport \ length}{Sum \ (reads \ mapped \ to \ transcript \ / \ transcript \ length)}
$$

(Transcript Per Million), a slight modification of Richards, and proposed (Li and Development Per Million)<br>TPM =  $10^6 * \frac{reads\ mapped\ to\ transcript\ /\ transient\ length}{Sum\ (reads\ mapped\ to\ transcript\ /\ transient\ length)}$ <br>TPM and RPKM are closely related. It is straightforward to convert ,<br>TPM ar<br>below. below.<br>
TPM =  $10^6 * \frac{RPKM}{Sum (RPKM)}$ <br>By definition, TPM and RPKM are proportional. However, TPM is unit-less, and it additionally fulfils the

$$
\text{TPM} = 10^6 * \frac{RPKM}{Sum (RPKM)}
$$

By defil<br>By defil<br>invariar l<br>i<br>t By definition, The finite in the proportional receiver, Therm and RPC, and it additionally fulfil and the proportional invariant average criterion. For a given RNA sample, if you were to sequence one million full length tr Intantivariant average criterion. For a given RNA sample, if you were to seep for a given gene or<br>transcripts, a TPM value represents the number of transcripts you would have seen for a given gene or<br>3  $\frac{1}{3}$  isoform. The average TPM is equal to 10^6 (1 million) divided by the number of annotated transcripts in<br>a given annotation, and thus is a constant. TPM is a better unit for RNA abundance since it respects the<br>invariance pr a given annotation, and thus is a constant. This is a better annotes into testimate conserver respects the propor<br>invariance property and is proportional to the average rmc, and thus adopted by the latest<br>computational alg computational algorithms for transcript quantification such as RSEM (Li and Dewey 2011), Kallisto (Bray<br>et al. 2016) and Salmon (Patro et al. 2017). Therefore, TPM will be used in the subsequent discussions<br>unless mentione et al. 2016) and Salmon (Patro et al. 2017). Therefore, TPM will be used in the subsequent discussions<br>unless mentioned otherwise, and examples will be given to illustrate how it can be misused.<br>Given the utility of RPKM a

et al. 2016) and Salmon (Patro et al. 2016) (Patro et al. 2017). The transfer also in the subsequent discussio<br>unless mentioned otherwise, and examples will be given to illustrate how it can be misused.<br>Given the utility o Given the utility of RPKM and TPM in comparing gene expression values within a samp<br>surprising that researchers would also seek to use the metrics for comparisons across p<br>datasets. While conceptually valid, this type of c (<br>(<br>( Surprising that researchers would also seek to use the metrics for comparisons across projects and<br>datasets. While conceptually valid, this type of cross-sample comparison can be problematic. As TPM<br>values are already norm datasets. While conceptually valid, this type of cross-sample comparison can be problematic. As TPM<br>values are already normalized, t is easy to assume they should be comparable across samples.<br>Unfortunately, this is not al values are already normalized, t is easy to assume they should be comparable across samples.<br>Unfortunately, this is not always true. In this review, we illustrate typical scenarios in which direct<br>comparison of RPKM and TP Unfortunately, this is not always true. In this review, we illustrate typical scenarios in which direct<br>comparison of RPKM and TPM across samples is problematic. To demonstrate, three public datasets<br>were downloaded from t Comparison of RPKM and TPM across samples is problematic. To demonstrate, three public datasets<br>were downloaded from the Sequence Read Achieve (SRA) and processed with Salmon (Patro et al. 2017)<br>using Gencode (Harrow et al were downloaded from the Sequence Read Achieve (SRA) and processed with Salmon (Patro et al. 2017)<br>using Gencode (Harrow et al. 2012) Release 29. The choices were based upon in-house evaluations of<br>isoform quantification a using Gencode (Harrow et al. 2012) Release 29. The choices were based upon in-house evaluations of<br>isoform quantification algorithms (Zhang et al. 2017) and different gene models (Zhao 2014; Zhao and<br>Zhang 2015). isoform quantification algorithms (Zhang et al. 2017) and different gene models (Zhao 2014; Zhao and

## Sample preparation protocol can greatly affect expression values

Ribosomal RNA (rRNA) is the most highly abundant component of total RNA isolated from animal or<br>human cells and tissues, comprising the majority (>80% to 90%) of the molecules in a total RNA sample Sample prepa<br>Ribosomal RN<br>human cells a :<br>|
|
| human cells and tissues, comprising the majority (>80% to 90%) of the molecules in a total RNA sample<br>(O'Neil et al. 2013; Fang and Akinci-Tolun 2016). To allow efficient transcript/gene detection, highly<br>abundant rRNAs mu (O'Neil et al. 2013; Fang and Akinci-Tolun 2016). To allow efficient transcript/gene detection, highly<br>abundant rRNAs must be removed from total RNA before sequencing. Standard approaches include<br>selection of polyadenylate (O'Neil et al. 2013; Fang and Akinci-Tolun 2016). To allow efficient transcript/gene detection, highly<br>abundant rRNAs must be removed from total RNA before sequencing. Standard approaches include<br>selection of polyadenylate abundant ranks inter the removed from term and term of experiming. Communic approaches included<br>selection of polyadenylated RNA (polyA) transcripts using oligo (dT) primers, or depletion of rRNAs<br>through hybridization capt selection mainly capture followed by magnetic bead separation. However, the polyA+ selection<br>and rRNA depletion methods each have their unique advantages and limitations. In principle, polyA+<br>selection mainly captures matu and rRNA depletion methods each have their unique advantages and limitations. In principle, polyA+<br>selection mainly captures mature mRNAs with polyA tails, whereas the rRNA depletion method can<br>sequence both mature and imm selection mainly captures mature mRNAs with polyA tails, whereas the rRNA depletion method can<br>sequence both mature and immature transcripts.<br>Both polyA+ selection and rRNA depletion were evaluated for gene quantification

Both polyA+ selection and rRNA depletion were evaluated for gene quantification in clinical RNA sequence an initial extermination changing to<br>Both polyA+ selection and rRNA depletion were<br>sequencing using human blood and colon tissue<br>prepared and sequenced using both protocols. A |
|
|
| Both polyanting using human blood and colon tissue samples (Zhao et al. 2018). The same samples were<br>prepared and sequenced using both protocols. All the raw sequencing reads were deposited into the<br>4 prepared and sequenced using both protocols. All the raw sequencing reads were deposited into the  $\frac{4}{3}$ prepared and sequenced using both protocols. All the raw sequencing reads were deposited into the

NCBI Sequence Read Archive under the accession number SRP127360. All sequenced transcripts were<br>broken down into five categories according to their annotated biotypes in Gencode (**Figure 1A**). For both<br>blood and colon samp blood and colon samples, the most abundant category with polyA+ selection was protein-coding genes,<br>whereas in the rRNA depletion protocol it was small RNAs. As shown in **Figure 1A**, the sequenced RNA<br>repertoires between t whereas in the rRNA depletion protocol it was small RNAs. As shown in **Figure 1A**, the sequenced RNA<br>repertoires between the polyA+ selection and rRNA depletion protocols are quite different. As a result<br>of the different s repertoires between the polyA+ selection and rRNA depletion protocols are quite different. As a result<br>of the different sample preparation protocols, the TPM values are not directly comparable, despite that<br>they are derive of the different sample preparation protocols, the TPM values are not directly comparable, despite that<br>they are derived from the same sample. In the blood sample (Figure 1B) sequenced by the polyA+<br>selection, the top thre they are derived from the same sample. In the blood sample (**Figure 1B)** sequenced by the polyA+<br>selection, the top three genes represent only 4.2% of transcripts (HBA2:1.5%, S100A9:1.4%, and<br>FTL:1.3%). In contrast, in the they are derived from the same sample. In the blood sample (Figure 1B) sequenced by the polyA+<br>selection, the top three genes represent only 4.2% of transcripts (HBA2:1.5%, S100A9:1.4%, and<br>FTL:1.3%). In contrast, in the r selection, the top three genes (RN7SL2:34.3%, RN7SL1:31.4%; and<br>FTL:1.3%). In contrast, in the rRNA depletion, the top three genes (RN7SL2:34.3%, RN7SL1:31.4%; and<br>RN7SK:9.3%) represent 75% of sequenced transcripts. As a r FIL:1.3.3%) represent 75% of sequenced transcripts. As a result, the expression levels of many other<br>genes are artificially deflated in the rRNA depletion sample. For the blood sample, the log2 ratio of TPM<br>values between genes are artificially deflated in the rRNA depletion sample. For the blood sample, the log2 ratio of TPM<br>values between polyA+ selection and rRNA depletion was calculated for individual genes. The<br>distribution of log2 rat genes are around any assume that the repeated in the repeated in the relation problem in the distribution of log2 ratio is depicted in **Figure 1C**, in which the mean values for protein-coding and small RNA genes are shown values between polyA+ selection is depicted in **Figure 1C**, in which the mean values for protein-coding and small<br>RNA genes are shown as dotted lines. For protein-coding genes, TPM values tend to be higher in the<br>polyA+ se distribution of log2 ratio is depicted in Figure 1c, in which the mean values for protein coding and small RNA<br>RNA genes are shown as dotted lines. For protein-coding genes, TPM values tend to be higher in the<br>polyA+ selec

## The different distribution of mRNAs across tissue types can mislead comparisons

polyA+ selection, while for small RNAs, the tendency is exactly opposite.<br>The different distribution of mRNAs across tissue types can mislead comparisons<br>Since different tissues express diverse RNA repertoires, TPM values The different distribution of mRNAs across tissue types can mislead cor<br>Since different tissues express diverse RNA repertoires, TPM values<br>considered directly comparable. To demonstrate this point, RNA-seq sant ד<br>2<br>1 Express different tissues in the since that repertoire, TMM values across direct tissue considered directly comparable. To demonstrate this point, RNA-seq samples corresponding to six tissue<br>types from the same subject GTE considered an external of the same subject GTEX-N7MS were downloaded from the Genotype-Tissue Expression (GTEx) project (Carithers and Moore 2015) and processed. The percentages of transcripts from mitochondria and the top (GTEx) project (Carithers and Moore 2015) and processed. The percentages of transcripts from mitochondria and the top three most abundant transcripts are shown in **Figure 2A**. An examination of blood and heart tissues make mitochondria and the top three most abundant transcripts are shown in **Figure 2A**. An examination of<br>blood and heart tissues makes the problem clear. In heart, 48.3% of sequenced transcripts are from<br>mitochondria, while in mitochondria and the top three most abundant transcripts are shown in Figure 2A. An examination of<br>blood and heart tissues makes the problem clear. In heart, 48.3% of sequenced transcripts are from<br>mitochondria, while in b mitochondria, while in blood this percentage drops to as low as 1.5%. Mitochondria generate most of<br>the cell's supply of adenosine triphosphate (ATP), used as a source of chemical energy, and play an<br>important role in the the cell's supply of adenosine triphosphate (ATP), used as a source of chemical energy, and play an<br>important role in the control of cell death in cardiac myocytes (Gustafsson and Gottlieb 2008). Thus, it is<br>not surprising the cell's supply of adenosine triphosphate (ATP), used as a source of chemical energy, and play an<br>important role in the control of cell death in cardiac myocytes (Gustafsson and Gottlieb 2008). Thus, it is<br>not surprising in the control of surprising to see that mitochondrial genes are actively transcribed and highly expressed in heart. In<br>heart, the top three highly expressed genes correspond to *MT-ATP6*, *MT-ATP8* and *MT-CO3*, and<br>repre heart, the top three highly expressed genes correspond to *MT-ATP6*, *MT-ATP8* and *MT-CO3*, and<br>represent a total of 17.4% of transcripts (Figure 2A). In blood, the top three genes (HBA2, HBB and<br>HBA1) constitute as high heart, the top three highly expressed genes correspond to MT-ATT6, MT-ATT6 and MT-CO3, and<br>represent a total of 17.4% of transcripts (Figure 2A). In blood, the top three genes (HBA2, HBB and<br>HBA1) constitute as high as 81. HBA1) constitute as high as 81.8% of sequenced transcripts. Considering the sequenced RNA repertoires<br>differ so dramatically, direct comparison of TPM values across tissues can be misleading. differ so dramatically, direct comparison of TPM values across tissues can be misleading.<br>
5 differ so dramatically, direct comparison of TPM values across tissues can be misleading.

The blood transcriptome in Figure 2A has a high complement of globin KIM that could potentially<br>saturate next-generation sequencing platforms, masking lower abundance transcripts. To circumvent<br>this issue, many commerciall saturate near generation sequencially available globin RNA reduction kits have been developed (Mastrokolias et<br>al. 2012; Shin et al. 2014). The top three genes (HBA2, HBB and HBA1) in this blood sample constitute as<br>high a al. 2012; Shin et al. 2014). The top three genes (HBA2, HBB and HBA1) in this blood sample constitute as<br>high as 81.8% of sequenced transcripts. If a very effective globin reduction kit is used, all goblins are<br>efficiently high as 81.8% of sequenced transcripts. If a very effective globin reduction kit is used, all goblins are<br>efficiently cleared. Accordingly, compared to RNA-seq without globin reduction, TPM values for the<br>remaining genes i hefficiently cleared. Accordingly, compared to RNA-seq without globin reduction, TPM values for the remaining genes in the same sample will increase about five-fold after globin reduction. This is another example where dif remaining genes in the same sample will increase about five-fold after globin reduction. This is another<br>example where differences in TPM values would be due to the experimental protocol and not<br>biologically relevant. remaining genes in the same sample will increase about for the experimental protocol and not<br>biologically relevant.<br>RNA compartmentalization affects TPM values between cytosolic and nuclear RNA-seq<br>The sterting metasial fo

#### RNA compartmentalization affects TPM values between cytosolic and nuclear RNA-seq

example where an example when the differences in the due to the experimental protocol and not<br>biologically relevant.<br>The starting material for RNA-seq studies is usually total RNA or polyA+ enriched RNA. Several<br>limitation **EXA Compartmentally**<br>The starting material<br>limitations arise from ||
|
|}<br>|} Imitations arise from analysing these heterogeneous pools of RNA molecules from nucleus, cytoplasm<br>and mitochondria. Although total RNA-seq has been shown to provide insight into ongoing transcription<br>and co-transcriptiona and mitochondria. Although total RNA-seq has been shown to provide insight into ongoing transcription<br>and co-transcriptional splicing in the nucleus (Tilgner et al. 2012), the simultaneous presence of mature<br>RNAs from the and co-transcriptional splicing in the nucleus (Tilgner et al. 2012), the simultaneous presence of mature<br>RNAs from the cytoplasm confounds the analysis of nuclear RNA maturation steps. Thus, the RNA-seq of<br>separated cytos RNAs from the cytoplasm confounds the analysis of nuclear RNA maturation steps. Thus, the RNA-seq of<br>separated cytosolic and nuclear RNA (Figure 2B) can significantly improve the analysis of complex<br>transcriptomes from mam separated cytosolic and nuclear RNA (Figure 2B) can significantly improve the analysis of complex<br>transcriptomes from mammalian tissues (Zaghlool et al. 2013). In comparison with conventional polyA+<br>RNA, cytoplasmic RNA co separated cytosolic and nuclear RNA (Figure 2B) can significantly improve the analysis of complex<br>transcriptomes from mammalian tissues (Zaghlool et al. 2013). In comparison with conventional polyA+<br>RNA, cytoplasmic RNA co The system of the manufacture of the system of the strange method in the strengtherm manufacture projections<br>sensitivity in expression analysis and splice junction detection. Conversely, the nuclear fraction shows an<br>enric REA, THE ENTRICAL CONTROLLATE CONTROLLATE CONTROLLATES (SPECIES) IS also are sensitivity in expression analysis and splice junction detection. Conversely, the nuclear fraction shows an enrichment of unprocessed RNA compare enrichment of unprocessed RNA compared with total RNA-seq, making it suitable for analysis of nascent<br>transcripts and RNA processing dynamics (Zaghlool et al. 2013). Considering the large differences in RNA<br>repertoires bet transcripts and RNA processing dynamics (Zaghlool et al. 2013). Considering the large differences in RNA<br>repertoires between nucleus and cytoplasm (Tilgner et al. 2012), the direct comparison of TPM values<br>across cellular transcripts and ANA processing a gradient congress of mattering are angles in the large differences repertoires between nucleus and cytoplasm (Tilgner et al. 2012), the direct comparison of TPM values<br>across cellular compa

## The "strandness" of RNA-seq has a substantial impact on transcriptome profiling

repertonce between and cytoplasm (MgM) of all 2012), the direct comparison of Tilm values<br>across cellular compartments of the same sample or between samples is not recommended.<br>The "strandness" of RNA-seq has a substantial The "strandness" of RNA-seq has a substantial impact on transcriptome profiling<br>Non-stranded RNA-seq does not retain the strand specificity of origin for each sequencing recorrent or between same samples is not retain the ך<br>|}<br>|} Strand information it is difficult - sometimes impossible - to accurately quantify expression levels for<br>genes with overlapping genomic loci that are transcribed from opposite strands (Pomaznoy et al. 2019).<br>In contrast, s strand information it is different interfactor in the accuration, quality, inpresent terms of genes with overlapping genomic loci that are transcribed from opposite strands (Pomaznoy et al. 2019).<br>In contrast, stranded RNA genes with overlapping genomic loci that are transcribed from opposite strand, and thus can resolve read<br>ambiguity in overlapping genes transcribed from opposite strands to provide a more accurate<br>quantification of gene ex In a ambiguity in overlapping genes transcribed from opposite strands to provide a more accurate<br>
a reading the stranded from opposite strands to provide a more accurate<br>
quantification of gene expression levels (Zhao et a quantification of gene expression levels (Zhao et al. 2015). The scatter plots of gene expression profiles  $\overline{6}$  $\epsilon$  for four biological replicates of blood samples (raw data downloaded from SRA under accession SRP056985) are shown in **Figure 3**. When comparing the same samples sequenced by the non-stranded and stranded protocols, there SRP 050559 are shown in Figure 3. When comparing the same samples sequenced by the non-stranded<br>and stranded protocols, there are many genes that are poorly correlated. It is not unusual that there are<br>genes whose expressi genes whose expression levels are high in one protocol, but very low or even zero in the other protocol.<br>When the stranded versus non-stranded sequencing groups were compared, as many as 1751 genes<br>were identified to be di When the stranded versus non-stranded sequencing groups were compared, as many as 1751 genes<br>were identified to be differentially expressed (a fold change greater than 1.5 and a Benjamini-Hochberg<br>adjusted p-value smaller When the stranded versus non-stranded verginating groups were semipately as many as 1751 genes<br>were identified to be differentially expressed (a fold change greater than 1.5 and a Benjamini-Hochberg<br>adjusted p-value smalle adjusted p-value smaller than 0.05) (Zhao et al. 2015). Thus, whether an analysis uses stranded RNA-seq<br>or not has a substantial impact on transcriptome profiling and expression measurements for many<br>genes. adjustive p-called and terms correlated provided by the stranded and expression measurements for many<br>genes.<br>Caution on RPKM and TPM comparison across samples with varying mRNA levels<br>RPKM and TPM regresent relative shunde

## Caution on RPKM and TPM comparison across samples with varying mRNA levels

genes.<br>Caution on RPKM and TPM comparison across samples with varying mRNA levels<br>RPKM and TPM represent relative abundance of a gene or transcript in a sample. The direct comparison<br>of RPKM and TPM across samples is meani g<br>Cautio<br>RPKM<br>of RPK ( F ( F RPKM and TPM across samples is meaningful only when there are equal total RNAs between<br>compared samples and the distribution of RNA populations are close to each other. Although equal total<br>RNAs are generally expected, it compared samples and the distribution of RNA populations are close to each other. Although equal total<br>RNAs are generally expected, it is rarely tested and not always met. For instance, cellular stress can<br>dramatically alt RNAs are generally expected, it is rarely tested and not always met. For instance, cellular stress can<br>dramatically alter the amount of RNA in cells, as shown for heat-shock treated cells (van de Peppel et al.<br>2003). Furth dramatically alter the amount of RNA in cells, as shown for heat-shock treated cells (van de Peppel et al.<br>2003). Furthermore, a comparison of embryonic stem cells and fibroblasts revealed a 5.5-fold difference<br>in mRNA lev 2003). Furthermore, a comparison of embryonic stem cells and fibroblasts revealed a 5.5-fold difference<br>in mRNA levels (Islam et al. 2011). Additionally, it was recently found that cells with high levels of c-Myc<br>can ampli 2013). Additionally, it was recently found that cells with high levels of c-Myc<br>2013). Furthermore, and amplify their gene expression program, producing two to three times more total RNA and<br>2013). Thus, under both<br>2013 ha in maniforcel (in the allegels) program, producing two to three times more total RNA and<br>generating cells that are larger than their low-Myc counterparts (Nie et al. 2012). Thus, under both<br>natural and experimental conditi can amplify their gene expression, producing the to three times more true times and a<br>generating cells that are larger than their low-Myc counterparts (Nie et al. 2012). Thus, under both<br>natural and experimental conditions natural and experimental conditions, the critical assumption that cells produce similar levels of RNA/cell<br>between cell types, disease states or developmental stages is not always valid. Depending on severity,<br>these differ between cell types, disease states or developmental stages is not always valid. Depending on severity,<br>these differences can influence the biological interpretation of gene expression values. RPKM and TPM<br>represent relativ between cell types, an influence the biological interpretation of gene expression values. RPKM and TPM represent relative abundance of transcripts in a sample but do not normalize for global shifts in total RNA contents (A represent relative abundance of transcripts in a sample but do not normalize for global shifts in total

### Discussions and Conclusions

represent relative abundance abundance of transcripts in a sequence of transcripts in a sequenced RNA repertoire can vary due to differences in RNA extraction & isolation protocols (total<br>The sequenced RNA repertoire can v **Discussions and Conclusions**<br>The sequenced RNA repertoire ca<br>RNA-seq vs polyA+ selection), dif  $\frac{1}{6}$ The sequence and RNA relation and RNA repeation of the set of the set of the set of the RNA repeation protocols (stranded vs non-stranded), and RNA abundance differences in mitochondrial and nuclear RNA compartments across and RNA-abundance differences in mitochondrial and nuclear RNA compartments across tissues. Such<br>differences should be controlled prior to comparing mRNA abundances across samples, even when using<br>7 differences should be controlled prior to comparing mRNA abundances across samples, even when using<br>7  $\frac{1}{\sqrt{2}}$ 

- The maximum is a supplies.<br>The M normalization. Below is a supplies a supplies are sequenced using the same protocol in terms of strandedness. If not,<br>Samples cannot be compared.<br>2. Make sure both samples use the same RNA
- 1. Make sumples.<br>
1. Make sumples.<br>
2. Make sum.<br>
2. Check t 1. Samples cannot be compared.<br>
1. Make sure both samples use the same RNA isolation approach (polyA+ selection vs ribosomal<br>
1. Make sure both samples use the same RNA isolation approach (polyA+ selection vs ribosomal<br>
1. samples cannot be compared.<br>
2. Make sure both samples use the same RNA isolation approach (polyA+ selection vs ribosomal<br>
RNA depletion). If not, they should not be compared.<br>
3. Check the fraction of the ribosomal, mitoc
- 2. Make sure both samples are and same random suppression (polyA+ selection vs random same<br>2. Check the fraction of the ribosomal, mitochondrial and globin RNAs, and the top highly<br>expressed transcripts and see whether suc Check the fraction of the ribosomal, mitochondrian<br>expressed transcripts and see whether such RNAs compared.<br>reads in a sample, and thus decrease the sequencing<br>enes in that sample. If the calculated fractions in expressed transcripts and see whether such RNAs constitute a very large part of the sequenced<br>reads in a sample, and thus decrease the sequencing 'real estate' available for the remaining<br>genes in that sample. If the calcu reads in a sample, and thus decrease the sequencing 'real estate' available for the remaining<br>genes in that sample. If the calculated fractions in two samples differ significantly, do not<br>compare RPKM or TPM values directl reads in a sample, and thus decrease the sequencing 'real estate' available for the remaining<br>genes in that sample. If the calculated fractions in two samples differ significantly, do not<br>compare RPKM or TPM values directl

genes in the compare RPKM or TPM values directly.<br>
Suite of the calculate samples when the total RNA contents<br>
distributions are very different. However, under appropriate circumstances, TPM can be still<br>
are suelitative c compare runners and taking an empty.<br>
Suid-never be used for quantitative condistributions are very different. Howe<br>
Suid-new suid-new partial and the suite of the suite |
|
|
|
| and its distributions are very different. However, under appropriate circumstances, TPM can be still<br>useful for qualitative comparison such as PCA and clustering analysis. In practice, it's not common to<br>use RPKM or TPM di useful for qualitative comparison such as PCA and clustering analysis. In practice, it's not common to<br>use RPKM or TPM directly in differential analysis. Instead, counts-based methods such as DESeq (Anders<br>and Huber 2010) use RPKM or TPM directly in differential analysis. Instead, counts-based methods such as DESeq (Anders<br>and Huber 2010) and edgeR (Robinson et al. 2010; Robinson and Oshlack 2010) have been developed to<br>identify differentia and Huber 2010) and edgeR (Robinson et al. 2010; Robinson and Oshlack 2010) have been developed to<br>identify differentially expressed (DE) genes. The fundamental assumptions underlying DESeq and edgeR<br>are summarized as foll identify differentially expressed (DE) genes. The fundamental assumptions underlying DESeq and edgeR are summarized as follows.<br>
1. Most genes are not DE.

- 
- 
- in the fundamental consequences (DE) generations in the fundamental assumptions underlying DESE and subsetsed as<br>are summarized as follows.<br>2. DE and non-DE genes behave similarly.<br>3. Balanced expression changes, i.e. the 1. Most genes are not<br>2. DE and non-DE genes<br>3. Balanced expression<br>are comparable. 3. Balanced expression changes, i.e. the number and magnitude of up- and down-regulated genes

2. DE and non-DE genes behave similarly.<br>
3. Balanced expression changes, i.e. the number and magnitude of up- and down- regulated genes<br>
are comparable.<br>
Normalization methods would perform poorly when the assumptions abo 3. Balanced expression changes, i.e. the number and magnitude of up- and down- regulated generalization methods would perform poorly when the assumptions above are violated. RNA-sequentlization plays a crucial role to ensu Exation methods<br>
Exation plays a cr<br>
(Dillies et al. 2013)<br>
malization methods normalization plays a crucial role to ensure the validity of gene counts for downstream differential<br>analysis (Dillies et al. 2013; Costa-Silva et al. 2017). However, to select the right between-sample RNA-<br>seq normalizati analysis (Dillies et al. 2013; Costa-Silva et al. 2017). However, to select the right between-sample RNA-<br>seq normalization methods for differential analysis is beyond the scope of this review, and reviewed<br>elsewhere (Evan seq normalization methods for differential analysis is beyond the scope of this review, and reviewed<br>elsewhere (Evans et al. 2018).<br>As more and more RNA-seq datasets are generated, meta-analyses of large-scale RNA-seq data

becoming increasingly common. In this review, we illustrated how easily RPKM and TPM can be elsewhere (Evans et al. 2021).<br>As more and more RNA-seq d<br>becoming increasingly comm<br>unintentionally misused, resul  $\frac{1}{1}$ becoming increasingly common. In this review, we illustrated how easily RPKM and TPM can be<br>unintentionally misused, resulting in misleading conclusions that can be attributed simply to technical<br>8 becoming in common increasing conclusions that can be attributed simply to technical<br>and TPM and TPM and TPM can be intributed simply to technical<br>and TPM can be attributed simply to technical<br>and TPM can be attributed sim  $8$  differences to which researchers may not be attuned. It can be reasonable to assume that the partitioning of total RNA among the different compartments (ribosomal RNA, pre-mRNA, mitochondrial RNA, genomic pre-mRNA and poly partitioning of total RNA and polyA+ RNA) of the transcriptome is comparable across samples in a given<br>RNA-seq project. This should be a key consideration in the initial experimental design. However, cross-<br>study analyses RNA-seq project. This should be a key consideration in the initial experimental design. However, cross-<br>study analyses are frequently done without proper control for these factors. Sequenced RNA repertoires<br>may change subs composition of the RNA population are close to each other when comparing RPKM/TPM values acrossstar, analyses are frequently done different experimental conditions and/or across different sequencing<br>protocols; thus, the proportions of gene expressions are not directly comparable in such cases.<br>Therefore, it is stron may protocols; thus, the proportions of gene expressions are not directly comparable in such cases.<br>Therefore, it is strongly recommended to always check whether the total RNA amount and the<br>composition of the RNA populati Therefore, it is strongly recommended to always check whether the total RNA amount and the composition of the RNA population are close to each other when comparing RPKM/TPM values across samples and sequenced RNA repertori Therefore, it is strongly recommended to always check when comparing RPKM/TPM values across<br>samples and sequenced RNA repertories. Otherwise, the comparison might be misleading, or become<br>even pointless. comples and sequenced RNA repertories. Otherwise, the comparison might be misleading, or become<br>even pointless.<br>**Competing financial interests** same pointless.<br>Samples and sequences are misleading RNA repeating financial interests.<br>All authors declare that they have no competing interests.

# even points of<br>Competing fin<br>All authors decl Competing financial interests

#### Author contributions

All authors declare that they have no competing interests.<br> **Author contributions**<br>
SZ conceived and designed the study. SZ, ZY and RS participated in writing the manuscript. All authors<br>
approved the final manuscript. 

### Acknowledgments

STE CONCERN Approved the final manuscript.<br>
STE conceived and Above and Above in writing the manuscript.<br>
The authors would like to thank Ken Dower, Enoch Huang and Abby Hill for their critical reading of the approved the final manuscript.<br>Acknowledgments<br>The authors would like to than<br>draft manuscript. The authors would be authors would be a strong and Abby Hill for the theory of the

#### References

- draft manuscript.<br> **References**<br>
Aanes H, Winata C, Moen LF, Ostrup O, Mathavan S, Collas P, Rognes T, Alestrom P. 2014. Normalization<br>
of RNA-sequencing data from samples with varying mRNA levels. *PLoS One* 9: e89158.  $\frac{1}{2}$
- 
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11.<br>
Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat<br>
Biotechnol 34: 52 of RNA-sequencing data from samples with varying mixtat levels. P205 One 9: e89158.<br>
S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Bill*<br>
Pimentel H, Melsted P, Pachter L. 2016. Near-o
- 
- Anders 3, Huber W. 2010. Differential expression analysis for sequence count data. Genome Bior 11.<br>Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification.<br>Biotechnol 34: 525-527. Bray NL, Finemer H, Melsted F, Fachter L. 2010. Near-optimal probabilistic RNA-seq quantification. Nut<br>Biotechnol 34: 525-527.<br>Carithers LJ, Moore HM. 2015. The Genotype-Tissue Expression (GTEx) Project. Biopreservation an biobanking 13: 307-308.<br>A, Madrigal P, Tarazona<br>DJ, Elo LL, Zhang X et al.<br>17: 13. biobanking 13: 307-308.<br>
Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney<br>
DJ, Elo LL, Zhang X et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Bio* Biobariking 13: 307-300.<br>A, Madrigal P, Tarazona !<br>DJ, Elo LL, Zhang X et al. 17: 13. Control P, Madrid P, Tarazona P, Tarazona P, McPherson A, McPherson A, Szczesniak MW, Szczesniak MW, Szczesnia<br>
17: 13.<br>
9  $DJ$ , Elo El, Zhang X et al. 2010. A survey of best practices for RNA-seq data analysis. Genome Biol **17**: 13. 17: 13.
- 
- The Silva of their solution of the Silva D, PLOS One 12: e0190152.<br>
Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D,<br>
Estelle J et al. 2013. A comprehensive evaluation o review and a software tool. *PLOS One* 12: e0190192.<br>The Rau A, Aubert J, Hennequet-Antier C, Jeanmougin<br>Estelle J et al. 2013. A comprehensive evaluation of<br>throughput RNA sequencing data analysis. *Briefings i*<br>, Hardin Estelle J et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-<br>throughput RNA sequencing data analysis. *Briefings in bioinformatics* 14: 671-683.<br>Evans C, Hardin J, Stoebel DM. 2018. Selecti
- 
- throughput NNA sequencing data analysis. Briefings in biomformatics 14: 071-083.<br>
Hardin J, Stoebel DM. 2018. Selecting between-sample RNA-Seq normalization is<br>
the perspective of their assumptions. *Briefings in bioinform*
- annotation and quantification using RNA-seq. Nat Methods 8: 469-477.
- Evans C, Hardin J, Akinci-Tolun R. 2016. Depletion of Ribosomal RNA Sequences from Single-Cell RNA-Sequencing<br>
Library. Current protocols in molecular biology 115: 7.27.21-27.22.20.<br>
Garber M, Grabherr MG, Guttman M, Trapn the perspective of their assumptions. *Briefings in biomformatics* 19. 776-792.<br>Akinci-Tolun R. 2016. Depletion of Ribosomal RNA Sequences from Single-Ce<br>Library. Current protocols in molecular biology 115: 7.27.21-27.27.2
- Elbrary. Current protocols in molecular biology 115: 7.27.21-27.21.20.<br>
Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome<br>
annotation and quantification using RNA-seq. Nat Methods Eibrary. Current protocols in molecular biology 115: 7.27.21-27.27.20.<br>M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational m<br>annotation and quantification using RNA-seq. Nat Methods 8: 469-477<br>son AB, Gottlieb RA. 2 annotation and quantification using RNA-seq. Nat Methods 8: 469-477.<br>Gustafsson AB, Gottlieb RA. 2008. Heart mitochondria: gates of life and death. Cardiovascular research<br>77: 334-343.<br>Harrow J, Frankish A, Gonzalez JM, Ta Gustafsson AB, Gottlieb RA. 2008. Heart mitochondria: gates of life and death. Cardiovascular research<br> **77**: 334-343.<br>
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A Gustafsson AB, Gottlieb RA. 2008. Heart influctioniana. gates of life and death. Cardiovascular research<br>
77: 334-343.<br>
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A,<br>Searle S. 2012. GENCODE: the reference human genome annotation for the ENCODE project.<br>Genome Res 22.<br>Islam S, Kjallqu Searle S. 2012. GENCODE: the reference human genome annotation for the ENCODE project.<br>
Genome Res 22.<br>
Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. 2011. Characterization of the<br>
single-c
- 
- 
- Genome Res 22.<br>Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. 2011. Characterization of the<br>single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 21: 1160-1167.<br>i Z, Figler B, Genome Res 22.<br>Kjallquist U, Mol<br>single-cell transc<br>1 Z, Figler B, Zha<br>discovery and de<br>wey CN. 2011. R<br>reference genom<br>0 Y, Shon HS, Ryu Islam S, Figler B, Zhang H, Cheng F. 2014. Introduction to RNA-Seq and its applications to drug<br>discovery and development. *Drug development research* 75: 324-330.<br>Li B, Dewey CN. 2011. RSEM: accurate transcript quantifica single-cell transcriptional randscape by highly multiplex RNA-seq. Genome Res 21: 1160-1167.<br>
1 Z, Figler B, Zhang H, Cheng F. 2014. Introduction to RNA-Seq and its applications to dr<br>
discovery and development. *Drug deve*
- 
- discovery and development. *Drug development research* 75: 324-330.<br>
Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a<br>
reference genome. *BMC Bioinformatics* 12: 323.<br>
Li P discovery and development: *Drug development research 75*: 324-330.<br>wey CN. 2011. RSEM: accurate transcript quantification from RNA-Se<br>reference genome. *BMC Bioinformatics* **12**: 323.<br>D Y, Shon HS, Ryu KH. 2015. Comparing reference genome. *BMC Bioinformatics* 12: 323.<br>Li P, Piao Y, Shon HS, Ryu KH. 2015. Comparing the normalization methods for the differential analysis of<br>Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16: 347. reference genome. BMC Biomformatics 12: 323.<br>
S Y, Shon HS, Ryu KH. 2015. Comparing the norm<br>
Illumina high-throughput RNA-Seq data. *BMC Bic*<br>
colias A, den Dunnen JT, van Ommen GB, t He<br>
sensitivity of next generation se Li P, Pian H, Pian H, Piao H, Piao M, Biology and States (16:347).<br>
Mastrokolias A, den Dunnen JT, van Ommen GB, t Hoen PA, van Roon-Mom WM. 2012. Increased<br>
sensitivity of next generation sequencing-based expression profi Illumina high-throughput RNA-Seq data. BMC Biomformatics 16: 347.<br>Iolias A, den Dunnen JT, van Ommen GB, t Hoen PA, van Roon-N<br>sensitivity of next generation sequencing-based expression profiling<br>human blood RNA. BMC genom
- 
- Substrivity of next generation sequencing-based expression profiling after globin reduction in<br>human blood RNA. *BMC genomics* 13: 28.<br>Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying m Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell<br>151: 68-79.<br>O'Neil D, Glowatz H, Schlumpberger M. 2013. Ribosomal RNA depletion for efficient use of RNA-seq Human blood RNA. *BMC genomics* 13: 28.<br>Ivi A, Williams BA, McCue K, Schaeffer L,<br>transcriptomes by RNA-Seq. *Nat Methods*<br>u G, Wei G, Cui K, Yamane A, Resch W, Wa<br>Myc is a universal amplifier of expressed<br>151: 68-79.<br>D, G Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R et al. 2012. c-<br>Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell<br>151: 68-79.<br>O'Neil D, G transcriptomes by RNA-seq. *Nut Wethods* 5: 621-626.<br>u G, Wei G, Cui K, Yamane A, Resch W, Wang R, Greer<br>Myc is a universal amplifier of expressed genes in lyr<br>**151**: 68-79.<br>p. Glowatz H, Schlumpberger M. 2013. Ribosomal R
- Myc is a universal ampliner of expressed genes in lymphocytes and embryonic stem cells. Cell<br>151: 68-79.<br>
0, Glowatz H, Schlumpberger M. 2013. Ribosomal RNA depletion for efficient use of RNA-seq<br>
capacity. Current protoco 151: 08-75.<br>
), Glowatz F<br>
capacity. *Cu*<br>
, Duggal G,<br>
quantificatic<br>
oy M, Seth<br>
estimates fr<br>
n MD, McCa
- capacity. Current protocols in molecular biology Chapter 4: Unit 4.15.<br>
Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon pro-<br>
quantification of transcript expression. Nat Methods 14.<br>
oy M, Sethi A, Greenbaum J,
- Pomaznoy M, Sethi A, Greenbaum J, Peters B. 2019. Identifying inaccuracies in gene expression quantification of transcript expression. Nat Methods 14.<br>
Pomaznoy M, Sethi A, Greenbaum J, Peters B. 2019. Identifying inaccuracies in gene expression<br>
estimates from unstranded RNA-seq data. Scientific reports 9: 16342.<br>
- 
- quantification of transcript expression. Nat Methods 14.<br>
Pomaznoy M, Sethi A, Greenbaum J, Peters B. 2019. Identifying inaccuracies in gene expression<br>
estimates from unstranded RNA-seq data. Scientific reports 9: 16342.<br> quantification of transcript expression. Nat Methods 14.<br>oy M, Sethi A, Greenbaum J, Peters B. 2019. Identi<br>estimates from unstranded RNA-seq data. *Scientific repo*<br>n MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Biocondu<br>ana estimates from unstranded RNA-seq data. Scientific reports 9: 16342.<br>
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression<br>
analysis of digital gene expression data. Bioinform
- estimates from unstranded RNA-seq data. Scientific reports 9. 10342.<br>
In MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package<br>
analysis of digital gene expression data. *Bioinformatics* 26.<br>
In MD, Oshlack A. 201 analysis of digital gene expression data. *Bioinformatics* 26.<br>Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of<br>RNA-seq data. *Genome Biol* 11: R25.<br>Shin H, Shannon CP, F analysis of digital gene expression data. *Biomformatics* 26.<br>
In MD, Oshlack A. 2010. A scaling normalization method 1<br>
RNA-seq data. *Genome Biol* 11: R25.<br>
Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilso<br>
Tebbu RNA-seq data. *Genome Biol* 11: R25.<br>
Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilson-McManus JE, Ng RT, McManus BM,<br>
Tebbutt SJ. 2014. Variation in RNA-Seq transcriptome profiles of peripheral whole bloo RNA-seq data. Genome Biol 11: R25.<br>Shannon CP, Fishbane N, Ruan J, Zho<br>Tebbutt SJ. 2014. Variation in RNA-thealthy individuals with and without<br>H, Knowles DG, Johnson R, Davis CA,<br>Guigo R. 2012. Deep sequencing of sco-tran
- Tebbutt SJ. 2014. Variation in RNA-Seq transcriptome profiles of peripheral whole blood from<br>healthy individuals with and without globin depletion. PLoS One 9: e91041.<br>Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabort The althy individuals with and without globin depletion. PLoS One 9: e91041.<br>H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA healthy marviduals with and without groom depiction. Thos One 9: e91041.<br>H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Sn<br>Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicin The Guigo R, 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22: 1616-1625.<br>10 Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly<br>co-transcriptional in the human genome but inefficient for IncRNAs. *Genome Res* 22: 1616-1625.<br>10 co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22: 1616-1625.
- van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC. 2003. Monitoring<br>global messenger RNA changes in externally controlled microarray experiments. *EMBO reports*<br>4:387-393.<br>Wagner GP, Kin K, Ly
- 
- 4: 387-393.<br>4: 387-393.<br>6P, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM<br>measure is inconsistent among samples. Theory Biosci 131: 281-285.<br>1 A, Ameur A, Nyberg L, Halvardson J, Grabherr M, 4: 387-333.<br>
GP, Kin K,<br>
measure is i<br>
I A, Ameur<br>
fractionatio<br>
tissues. *BM*<br>
C, Dower K,<br>
validation o measure is inconsistent among samples. Theory Biosci 131: 281-285.<br>
1 A, Ameur A, Nyberg L, Halvardson J, Grabherr M, Cavelier L, Feu<br>
fractionation improves RNA sequencing analysis of mature and nasc<br>
tissues. *BMC biotec*
- measure is inconsistent among samples. Theory Biosci 131: 281-285.<br>
Zaghlool A, Ameur A, Nyberg L, Halvardson J, Grabherr M, Cavelier L, Feuk L. 2013. Efficient cellular<br>
fractionation improves RNA sequencing analysis of m Engine Line (minimal or process RNA sequencing analysis of mature and nascent transcripts from human<br>tissues. *BMC biotechnology* 13: 99.<br>Zhang C, Dower K, Zhang B, Martinez RV, Lin LL, Zhao S. 2018. Computational identifi fissues. *BMC biotechnology* 13: 99.<br>
2, Dower K, Zhang B, Martinez RV, Lin LL, Zhao S. 2018. Computational identification and<br>
validation of alternative splicing in ZSF1 rat RNA-seq data, a preclinical model for type 2 di tissues. *BMC biotechnology* 13: 99.<br>
2, Dower K, Zhang B, Martinez R<br>
validation of alternative splicing in .<br>
nephropathy. *Scientific reports* **8**: 7<br>
7, Zhang B, Lin L-L, Zhao S. 2017. Ev.<br>
isoform quantification. *BMC* validation of alternative splicing in ZSF1 rat RNA-seq data, a preclinical model for type 2 diabetic<br>nephropathy. Scientific reports 8: 7624.<br>Zhang C, Zhang B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computatio
- nephropathy. Scientific reports 8: 7624.<br>
7 Zhang B, Lin L-L, Zhao S. 2017. Evaluat<br>
isoform quantification. *BMC genomics* 1<br>
2014. Assessment of the impact of usir<br>
reads. *PLoS One* 9: e101374.<br>
Fung-Leung WP, Bittner A
- valid properties of alternative splitch of alternative splitch of alternative splitch of alternative splitch splicing B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computational tools for RNA-seq isoform quantific Example, 2011, 201 Soform quantification. *BMC genomics* 18: 583.<br>2014. Assessment of the impact of using a ref<br>reads. *PLoS One* 9: e101374.<br>Fung-Leung WP, Bittner A, Ngo K, Liu X. 20<br>transcriptome profiling of activated T cells. *PLo*:<br>Zha
- Thao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. 2014. Comparison of RNA-Seq and microarray in<br>
transcriptome profiling of activated T cells. *PLoS One* 9: e78644.<br>
Zhao S, Zhang B. 2015. A comprehensive evaluation of ensem
- 
- reads. PLOS One 9: e101374.<br>Fung-Leung WP, Bittner A,<br>transcriptome profiling of act<br>Zhang B. 2015. A compreher<br>context of RNA-seq read map<br>Zhang B, Gordon W, Zhang Y,<br>RNA-Seq data analysis. In B<br>InTech. The System of Function of the Single of Tells. PLoS One 9: e78644.<br>
Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the<br>
context of RNA-seq read mapping and gene quantification Thang B. 2015. A comprehensive evaluation of ensembl, RefSe<br>context of RNA-seq read mapping and gene quantification. *BMC*<br>Zhang B, Gordon W, Zhang Y, Du S, Paradis T, Vincent M, Von Sc<br>RNA-Seq data analysis. In *Bioinform*
- Chao S, Zhang B, Gordon W, Zhang Y, Du S, Paradis T, Vincent M, Von Schack D. 2016. Bioinformatics for<br>RNA-Seq data analysis. In *Bioinformatics Updated Features and Applications*, pp. 125-149,<br>InTech.<br>Zhao S, Zhang Y, G context of RNA-seq read mapping and gene quantification. *BMC genomics* 16: 97.<br>Zhang B, Gordon W, Zhang Y, Du S, Paradis T, Vincent M, Von Schack D. 2016. Bio<br>RNA-Seq data analysis. In *Bioinformatics - Updated Features a* 2013, 2014 Separation Marketines S, 2014 Separation S, 2014 and S, 2014 and S, Zhang Y, Gamini R, Zhang B, von Schack D. 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: p RNA-Seq data analysis. In Biomformatics - Opdated Features and Applications, pp. 125-149,<br>InTech.<br>Zhang Y, Gamini R, Zhang B, von Schack D. 2018. Evaluation of two main RNA-seq approaches for<br>gene quantification in clinica There is<br>Zhang Y,<br>gene qu<br>reports<br>Zhang Y<br>and nongenomid
- Entrary Entrary Controllers, 2018. Comparison in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Scientific<br>reports 8: 4781.<br>Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, von Schack D, Zhang B. 2015. Comp gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion: Scientific<br>reports 8: 4781.<br>Zhang Y, Gordon W, Quan J, Xi H, Du S, von Schack D, Zhang B. 2015. Comparison of stranded<br>and non-strande reports **8**: 4781.<br>Zhang Y, Gordor<br>and non-strand<br>genomics **16**: 67 and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC<br>genomics 16: 675.<br>The S, Chang B. 2015. Comparison of strandard B. 2015. Comparison of strandard B. 2015. and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC<br>genomics 16: 675. genomics 16: 675.



Figure 1. Comparison of TPM values of blood or colon samples with either polyA+ selection or rRNA A deletion. The same blood and colon RNA samples were sequenced by both protocols (denoted as polyA+ + and rRNA, respectively). A) The breakdown or sequenced transcripts by their biotype, b) The percentages of the top three highly expressed genes; and C) The distribution of log2 ratio of TPM values s in polyA+ selection over rRNA deletion.



Figure 2. A) The percentages of transcripts from mitochondria, and the top three most abundant t transcripts, in different tissue samples of the same subject (GTEX-N7MS) from the GTEx project. B) In n cellular fractionation RNA sequencing, the nucleic and cytosolic RNA populations are very different, and d thus TPM values are not directly comparable.



Figure 3. Scatter plots of gene expression profiles between stranded and non-stranded RNA-seq. For r blood biological replicates PFE1, PFE2, PFE3, and PFE4, the scattering patterns are consistent. While the e majority of genes are arrayed along the diagonal lines, there are still many genes whose expression n levels are dramatically impacted by sequencing protocols. The x- and y-axis represent Log2(RPKM).



## **samples and sequencing protocols Misuse of RPKM or TPM normalization when comparing across**

Shanrong Zhao, Zhan Ye and Robert Stanton

RNA published online April 13, 2020



identifier (DOIs) and date of initial publication. PubMed from initial publication. Citations to Advance online articles must include the digital object publication). Advance online articles are citable and establish publication priority; they are indexed by appeared in the paper journal (edited, typeset versions may be posted when available prior to final Advance online articles have been peer reviewed and accepted for publication but have not yet

**<http://rnajournal.cshlp.org/subscriptions>** To subscribe to RNA go to: